Module 2 Summary
group 7

Introduction

It is hard to see how much body fat a person has at a glance. Although there are many ways to estimate body fat, most of them are either overly sophisticated or lack accuracy. In this project, we are dedicated to come up with an effective model to predict human body fat percentage using some of the basic body measurements.

Data Cleaning and Preview

Our dataset contains body fat percentages and body density(from underwater weighing) of 252 men and their basic body measurements such as Age, Weight, Height and BMI. The BodyFat variable has a mean of 18.94% and standard deviation of 7.75. The average age, height and weight are 44.89, 178.92 pounds and 70.15 inches respectively.

Before we built our models, we made some modifications to the dataset. First, We changed the units of weight and height to kg and cm because all of the other variables in the dataset are measured in cm. Then, we fixed a person who is only 74.94 cm tall using the relationship between bmi, weight and height (bmi=kg/m^2). Last but not least, We removed the person with 0% body fat percentage, because this is scientifically impossible and we couldn't manage to fix this.

Model Interpretation

Our final model is:

$$BODYFAT = \alpha + \beta_1 weight + \beta_2 abdomen + \beta_3 weight * abdomen + \varepsilon$$

We tested our model by estimating the body fat percentage of Michael Phelps. His weight, height and abdomen circumference are 87 kg and 86.36 cm respectively. Based on these values, Michael's estimated bodyfat percentage is 10.50%. The 95% prediction interval for his body fat percentage is between 3.80% and 19.93%. This result is pretty close to what we got from Google, which is 10.3%.

Our estimated weight coefficient is -0.008, weight and abdomen interaction coefficient is -0.003, abdomen coefficient is 1.14 and intercept is -64.42. This means for a man who has 89 cm abdomen circumference, our model predicts that his body fat percentage will increase 0.27% for every 1 kg increase in his weight.

Model Selection

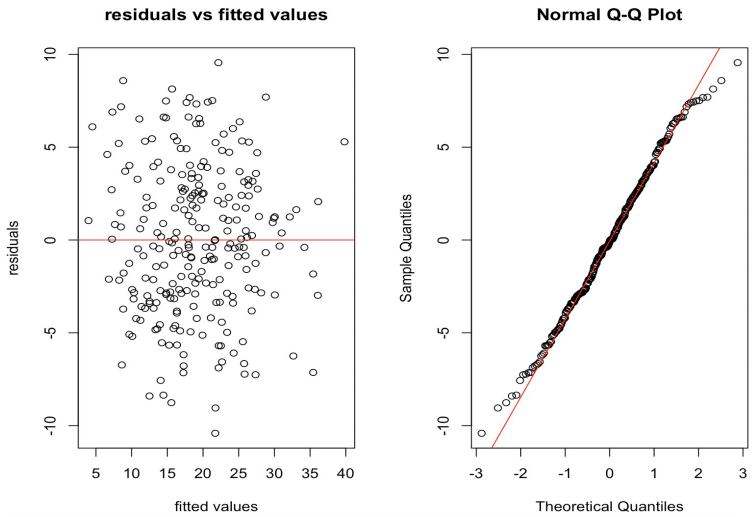
We tried three different models, stepwise linear regression model without interaction, linear regression with interaction and lasso. The adjusted R^2 are respectively 0.74, 0.72, 0.73, and MSE are 15.12, 16.08, 15.67 respectively. So among them, we select the linear regression model with interaction because they have relatively close R^2 and MSE, while our final model is the simplest one and can be easily measured.

We conducted a t-test to test the significance of our chosen predictors. Under Type-I error=0.05, we can conclude that abdomen and weight*abdomen both have p-value<0.05 and hence significant. The detailed result is shown in the table below:

predictor	coefficient	p-value	95% lower CI	95% upper CI
intercept	-64.42	4.63e-15	-79.60	-49.25
weight	-0.008	0.94	-0.21	0.19
abdomen	1.14	<2e-16	0.97	1.32
weight*abdomen	-0.003	0.001	-0.005	-0.001

Model Diagnostics

We test the normality of our residuals. From the normal Q-Q plot we can see that residuals are approximately normally distributed. And from residuals vs fitted values plot, we can see that residuals show no obvious pattern with mean zero and randomly distributed. So we believe the normality assumption is plausible. However from the plot, we can see that there are some obvious outliers. In conclusion, our model is correct and effective in general cases.



Our model is simple and easy to explain. Also the residuals of our model are approximately normally distributed and show no obvious pattern, indicating the effectiveness of our model. Weakness of our model is that we only use two predictors and thus lose some accuracy of the model. And the abdomen and weight data we have is rounded, which also introduces bias in prediction.

Conclusion

In our model, we first do some data cleaning to the extreme data. Then, compared with other models, we use the relatively simple multilinear regression model to predict density. Abdomen have significant positive effects on the response, and abdomen and weight interaction negatively influences body fat.

Contribution

1. Tongyue Jia wrote the introduction and conclusion in report and built the framework of shiny app and wrote most of the codes. She will present P15-17 in our slides.

2. Keyu Hu wrote the data cleaning and preview, model interpretation part in report. He will also present P1-8 of our slides.

3. Qilu Zhou wrote the model selection and model diagnostics part in the report. She modified the model used in shiny app and she will present P9-14 in the slides.